Available online at www.sciencedirect.com

ScienceDirect

Gene 415 (2008) 13–22

GENE

www.elsevier.com/locate/gene

A novel SINE family occurs frequently in both genomic DNA and transcribed sequences in ixodid ticks of the arthropod sub-phylum Chelicerata

Jack D. Sunter [a,1], Sonal P. Patel [a,1], Robert A. Skilton [a,e], Naftaly Githaka [a], Donald P. Knowles [b,c], Glen A. Scoles [b], Vishvanath Nene [d], Etienne de Villiers [a], Richard P. Bishop [a,*]

[a] The International Livestock Research Institute (ILRI), PO Box 30709, Nairobi, Kenya
[b] USDA ARS Animal Disease Research Unit, Washington State University, Pullman, WA 99164, USA
[c] Program in Vector-borne Diseases, Department of Veterinary Microbiology and Pathology, Washington State University, Pullman, USA
[d] Institute for Genome Sciences and Department of Microbiology and Immunology, University of Maryland School of Medicine HSF-II, S-447 20 Penn Street Baltimore, MD 21201, USA
[e] Biosciences eastern and central Africa (BecA), PO Box 30709, Nairobi, Kenya

Received 1 August 2007; received in revised form 14 January 2008; accepted 30 January 2008
Available online 21 February 2008
Received by J.V. Moran

Abstract

Reassociation kinetics and flow cytometry data indicate that ixodid tick genomes are large, relative to most arthropods, containing $\geq 10^9$ base pairs. The molecular basis for this is unknown. We have identified a novel small interspersed element with features of a tRNA-derived SINE, designated Ruka, in genomic sequences of *Rhipicephalus appendiculatus* and *Boophilus (Rhipicephalus) microplus* ticks. The SINE was also identified in expressed sequence tag (EST) databases derived from several tissues in four species of ixodid ticks, namely *R. appendiculatus*, *B. (R.) microplus*, *Amblyomma variegatum* and also the more distantly related *Ixodes scapularis*. Secondary structure predictions indicated that Ruka could adopt a tRNA structure that was, atypically, most similar to a serine tRNA. By extrapolation the frequency of occurrence in the randomly selected BAC clone sequences is consistent with approximately 65,000 copies of Ruka in the *R. appendiculatus* genome. Real time PCR analyses on genomic DNA indicate copy numbers for specific Ruka subsets between 5800 and 38,000. Several putative conserved Ruka insertion sites were identified in EST sequences of three ixodid tick species based on the flanking sequences associated with the SINEs, indicating that some Ruka transpositions probably occurred prior to speciation within the metastriate division of the Ixodidae. The data strongly suggest that Class I transposable elements form a significant component of tick genomes and may partially account for the large genome sizes observed.

Keywords: *Rhipicephalus*; *Boophilus*; *Ixodes*; Short interspersed repetitive element; tRNA-derived; Retroposon

1. Introduction

Short interspersed repetitive elements (SINEs), containing RNA polymerase III promoters, are major components of eukaryotic genomes that are particularly abundant in the heterochromatic compartment of vertebrates and plants (reviewed by Kidwell, 2002). SINEs are transposable elements whose ability to move to new locations is based on reverse transcription prior to genomic integration. However their

Abbreviations: AvGI, *Amblyomma variegatum* gene index; BAC, Bacterial artificial chromosome; BmiGI, *Boophilus* (*R.*) *microplus* gene index; EST, Expressed sequence tag; DFCI, Dana Faber Cancer Institute; G6PDH, Glucose-6-phosphate dehydrogenase; IsGI, *Ixodes scapularis* gene index; RaGI, *Rhipicephalus appendiculatus* gene index; bp, base pair(s); kb, kilo base(s); kpb, kilo base pairs; Ra BAC, *R. appendiculatus* BAC-derived contigs; SINEs, Short interspersed repetitive elements; LINE, Long interspersed element; TRAP, Tandem repeats analysis program.

[*] Corresponding author. Tel.: +254 20 4223000.
E-mail address: r.bishop@cgiar.org (R.P. Bishop).

[1] These authors contributed equally to this work.

0378-1119/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.01.026

transposition is non-autonomous since these elements do not encode the proteins required for mobility. Most SINEs are derived from tRNA (Okada, 1991), although some, such as the *Alu* family which accounts for approximately 10% of the human genome, are thought to originate from 7SL RNA sequences (Ullu and Tschudi, 1984). Recently it has been recognised that, due to the irreversible, independent nature of their insertion, SINEs represent a powerful molecular tool for defining phylogeny (reviewed by Shedlock and Okada, 2000). The occurrence of SINEs in insect genomes is variable. They are apparently absent from the euchromatic section of the relatively compact 1.8×10^8 bp genome of *Drosophila melanogaster* (Kaminker et al., 2002), but several distinct families of SINEs with copy numbers of up to 5.9×10^4 bp per genome have been described in *Ædes æygpti*, the mosquito vector of the yellow fever virus (Tu, 1999; Tu et al., 2004).

Ticks are arthropods in the sub-phylum Chelicerata (order *Acari*). Although the Ixodid and Argasid ticks are not well represented in the fossil record, representatives of the Acari are present in Devonian sediments (Klompen et al., 1996) indicating that the Acarinid lineage has been evolutionarily separated from the Insecta for a minimum of 350 million years. Ticks are important vectors of pathogens, including Lyme disease in humans and a range of protozoan and rickettsial diseases of livestock (Jongejan and Uilenberg, 2004). Tick-borne diseases are responsible for more than an estimated 17 billion dollars of economic loss in the livestock sector globally (De Castro, 1997).

The genome sizes of three species of ixodid ticks, *Amblyomma americanum* (Palmer et al., 1994), *Boophilus (Rhipicephalus) microplus* and *Ixodes scapularis* (Ullmann et al., 2005) have been estimated using reassociation kinetics. These studies indicated large genome sizes, at the upper end of the spectrum described for arthropod genomes, ranging from 1×10^9–7×10^9 bp, with a comparatively high content of moderately repetitive DNA, 38–42%. Seven species of ixodid ticks including *A. americanum* and *I. scapularis* have also had their genome sizes estimated by flow cytometry with predicted ranges from 1.8×10^9–3.2×10^9 bp (Geraci et al., 2007). Based on the sizes of other tick genomes, by extrapolation it is reasonable to presume that the *R. appendiculatus* genome is at least 1×10^9 bp. One group of insects with a comparable genome size (1.91×10^9) is Hawaiian crickets (*Laupala cerasina*) for which evidence exists that a low rate of loss of retroposon-derived sequences that have lost their ability to move is a key determinant of genome size (Petrov et al., 2000).

In order to underpin new approaches to understand the molecular basis of disease transmission in ticks and initiate research on arthropod comparative genomics, expressed sequence tag (EST) databases have been assembled and annotated as Dana Faber Cancer Institute (DFCI) Gene Indices (Quackenbush et al., 2001) for four ixodid tick species, *A. variegatum* (Nene et al., 2002), *R. appendiculatus* (Nene et al., 2004), *B. (R.) microplus* (Guerrero et al., 2005) and *I. scapularis* (Ribeiro et al., 2006). Here we describe construction and initial sample sequencing of a bacterial artificial chromosome (BAC) library of *R. appendiculatus* (Ra BAC) and the comparison of this data with the transcripts sequences in gene indices. Analysis of repetitive sequences within genomic and transcribed DNA reveals the widespread presence of a tRNA-derived SINE, that we have designated Ruka (from the Kiswahili verb ‘to jump’), in ixodid ticks.

2. Materials and methods

2.1. Preparation of Rhipicephalus appendiculatus high molecular weight DNA

Briefly, snap frozen *R. appendiculatus* eggs of the Muguga stock that has been maintained at ILRI for over 40 years were ground and resuspended in PBS (phosphate-buffered saline). After two washes in PBS, the material was resuspended in PBS and mixed with an equal volume of 1% low melting point agarose. Plugs were prepared from the cell/agarose mixture using a DNA plug mold kit (Bio-Rad). The plugs were incubated in 10 mM Tris–HCl/0.5 M EDTA (TE buffer), containing 1% *N*-lauryl sarcosine and 0.2% proteinase K at 50 °C overnight. Following dialysis in TE, the plugs were stored in TE at 4 °C.

2.2. Construction, initial characterisation and sequencing of random clones from a bacterial artificial chromosome (BAC) library

A BAC library was constructed in the pECBAC1 vector by Amplicon Express (Pullman WA, USA) from agarose plugs containing high molecular weight DNA from *R. appendiculatus*. Briefly, the DNA, partially digested with *Mbo*1 within the agarose, was ligated into the *Bam*H1 site of pECBAC1. Ligations were transformed into DH10b *E. coli* cells, and individual colonies were picked robotically and arrayed into 384 well plates. The library clones have an average insert size of 115 kbp, and based on a presumed genome size of 1×10^9 bp the library represents 5.5× coverage of the genome (data not shown). Three clones were randomly selected from the BAC library for nucleotide sequence determination. DNA sequencing and subsequent assembly was performed at TIGR as described (Desjardins et al., 2007).

2.3. Quantification of Ruka copy number in R. appendiculatus genomic DNA using quantitative real time PCR (RT-PCR)

2.3.1. Extraction of R. appendiculatus gDNA

Genomic DNA was extracted from *R. appendiculatus* eggs using the phenol-chloroform method described by Sambrook et al., 1989.

2.3.2. Construction of calibration curves for Ruka sequences

An *E. coli* plasmid clone identified within the *R. appendiculatus* gene index (Nene et al., 2004), RAAAQ49TF, containing a single copy of Ruka, was obtained from the *R. appendiculatus* salivary gland EST library (Nene et al., 2004). This was used as a standard for the RT-PCR reaction. Plasmid was prepared using the Wizard Plus SV Minipreps DNA Purification System (Promega) according to the manufacturer's instructions. The purified plasmid DNA was linearised with the restriction enzyme *Not*I. Linearisation was confirmed by agarose gel electrophoresis on a 1.2% Seakem preparative gel. The linearised plasmid DNA was gel purified using a QIAquick Purification Kit (Qiagen) according to the manufacturer's instructions. Plasmid DNA concentrations were measured using a Nanodrop spectrophotometer (NanoDrop, Wilmington, DE) and the copy numbers were calculated using the following equation:

$$\text{Copies}/\mu\text{l} = \frac{(6.02 \times 10^{23}\ \text{copies}) \times (\text{plasmid concentration g}/\mu\text{l})}{(\text{Number of bases}) \times (660\ \text{Da/base})}$$

Five-fold serial dilutions of the plasmid (10^7 copies to 10^3 copies) in nuclease-free water were prepared. Single use aliquots of the standard dilutions were stored at −80 °C to ensure plasmid DNA stability (Applied Biosystems, 2003; Godornes et al. 2007). Genomic DNA standards were made by serial dilution of *R. appendiculatus* egg genomic DNA (2.3–23000 pg).

2.3.3. Quantitative RT-PCR

Real time PCR was performed in triplicate with SYBR® GREEN PCR Master Mix (Applied Biosystems) with 300–500 nM primers, in a final volume of 25 µl. PCR was performed using a 7500 Real Time PCR System (Applied Biosystems) for 40 cycles at 95 °C for 10 s, 60 °C for 1 min, and 72 °C for 2 min. Primer sequences that were conserved among several different Ruka elements were derived from an alignment of the eight most conserved Ruka sequences present in the sequenced Ra BAC clones. The primer sequences are listed in Table 1. They were sorted into six pairs FWD_1 and REV_1 (Ruka 1), FWD_2 and REV_1 (Ruka 2), FWD_3 and REV_1 (Ruka 3), FWD_1 and REV_2 (Ruka 4), FWD_2 and REV_2 (Ruka 5) and FWD_3 and REV_2 (Ruka 6). A melting curve analysis was performed after the amplification phase, to check for non-specific amplification or primer–dimer formation.

The threshold cycle (Ct), the cycle number at which the fluorescence of the sample exceeded that of the background, was determined by 7500 Real Time PCR System sequence detection system version 1.2.1 (Applied Biosystems) using the standard curve method. Data analysis was performed using the same software.

2.4. In silico analyses

2.4.1. Sequence similarity searches

Sequences with similarity to those present in the BAC clones were identified within the non-redundant nucleotide and protein databases in GenBank using BLASTN and TBLASTX (Altschul et al., 1990). The *R. appendiculatus* BAC sequences were also searched against themselves. Sequences similar to Ruka (those with characteristics of tRNA-derived SINEs) were sought in the sequences from BAC clones by locally aligning the BAC sequences with Ruka using the BLASTN algorithm (Smith and Waterman, 1981) with cut-off values of $E \leq 1\text{e}^{-10}$ and sequence alignment length ≥180 nucleotides (rest of the default settings: low complexity filter ON; Matrix: BLOSUM62; Nucleotide mismatch penalty: −3; Nucleotide match reward: 1; Gap open cost: 5; Gap extension cost: 2).

Table 1
Primers used for quantitative real time PCR of Ruka from genomic DNA

Primer name	Sequence
FWD_1	5′ GYGGTTABGGBGCTCGRCTGCTGACC 3′
FWD_2	5′ GGBGCTCGRCTGCTGACCSGMAGGT 3′
FWD_3	5′ TCGRCTGCTGACCSGMAGGTHGCG 3′
REV_1	5′ TTATGAGRGACGCCGTAGTGGAGGGCT 3′
REV_2	5′ TGGGGTTTWACGTCCCAAAACCAC 3′

Publicly available databases of expressed sequence tags (ESTs) for several animal, plant and fungal species have been catalogued by the Dana Faber Cancer Institute (DFCI) Gene Index Project into species specific databases, designated as the Gene Indices (Quackenbush et al., 2000, 2001; The Gene Index Databases). Amongst these, gene indices for four tick species are available, namely *R. appendiculatus*, *B. (R.) microplus*, *A. variegatum* and *I. scapularis*. They are designated RaGI, BmiGI, AvGI and IsGI respectively. BLASTN was used to find sequences similar to Ruka in all four tick Gene Indices using cut-off values of $E \leq 1\text{e}^{-10}$ and sequence alignment length ≥180 bases (low complexity filter disabled). Sequences meeting these criteria were retrieved.

2.4.2. Repeat identification

Automated as well as manual approaches for repeat finding were performed. The latter method used local pairwise BLASTN (Smith and Waterman, 1981) as well as the Dot plot tool — Dotlet (Junier and Pagni, 2000) to identify and verify the repeat regions respectively. Tandem repeats were identified using Tandem Repeat Finder (Benson, 1999) along with Tandem Repeats Analysis Program (TRAP) (Sobreira et al., 2006). REPFIND (Betley et al., 2002) helped identify 100% identical direct repeats.

2.4.3. Sequence clustering

34 sequences from RaGI, 38 from BmiGI and 8 from the BAC sequences, that were similar to Ruka, were aligned using the alignment program CLUSTAL W (v. 1.81) (Chenna et al., 2003) with default settings (Gap open: 15, gap extension: 6.66, Matrix: DNA identity matrix). Each alignment was screened for obvious alignment errors and manually edited accordingly using JALVIEW — a java-based multiple-alignment editor (Clamp et al., 2004). The PHYLIP 3.6 package (Felsenstein, 1989) was used for the subsequent phylogeny analyses. One hundred bootstrap datasets (default parameters, random number seed = 5) were computed using DNAPARS — parsimony method for DNA data. DNAML, the maximum likelihood algorithm for DNA data, was used with default parameters to construct and display unrooted phylogenetic trees with the 80 sequences.

2.4.4. tRNA secondary structure prediction

tRNAscan-SE 1.21 (Lowe and Eddy, 1997) which was accessed via http://lowelab.ucsc.edu/tRNAscan-SE/ was used to identify a transfer RNA from a nucleotide sequence. The default

Table 2
Frequency of occurrence of Ruka-like sequences within transcript databases (RaGI, BmiGI, AvGI, IsGI) of four ixodid tick species, and BAC sequences of *R. appendiculatus*

Species	No. homologs with alignment length ≥180 bp	Alignment length range	% residue identity range
BmiGI	37	181–244	66–93
RaGI	34	180–242	64–86
R. appendiculatus BAC	8	189–231	84–87
AvGI	7	199–251	64–78
IsGI	3	208–242	68–70
Species	No. homologs with alignment length <180 bp	Alignment length range	% residue identity range
BmiGI	23	80–168	66–95
RaGI	16	99–179	70–91
R. appendiculatus BAC	20	53–172	83–96
AvGI	3	106–171	71–84
IsGI	0	n/a	n/a

search mode on eukaryotic source was used. It was also used for tRNA secondary structure prediction of Ruka.

3. Results

3.1. A novel tRNA-derived transposable element — Ruka

In order to provide initial insight into the organisation of tick genomic DNA three randomly chosen BAC clones from an *R. appendiculatus* BAC (Ra BAC) library were sequenced. Assembly of the BAC sequences was not completed most likely due to presence of repetitive sequences, but insufficient coverage of shotgun sequencing could also have been a contributing factor. The assembly program generated 12 contigs ranging from 2 kb–85 kb in length. The contigs were searched against each other using BLASTN (E-value ≤ $1e^{-10}$; Alignment length ≥ 180 bp) to locate repetitive elements present within them. Twenty eight copies of a conserved sequence approximately 240 bases in length were detected within the Ra BAC sequences (Table 2). The distribution of this repeat within the four longest contig assemblies (58–85 kbp) that were derived from three different BAC clones appeared to be random (Fig. 1). These repeat sequences, which were designated Ruka, comprise approximately 1.6% of the 250 kb of *R. appendiculatus* genomic sequence represented in the BAC contigs. Assuming (1) that these Ra BAC contigs are representative, and (2) a genome size for *R. appendiculatus* of 1×10^9, then a total of 65,000 copies of Ruka would be predicted.

One of the Ruka sequences from Ra BAC was searched against the GenBank database and was found to be most similar (E-value ≤ $1e^{-35}$; Identity ≥ 85%) to a genomic sequence within intron 7 of the glucose-6-phosphate dehydrogenase (G6PDH) gene from *B. (R.) microplus* (Accession no. DQ118973). The version of Ruka found in this intron was used as a reference in subsequent analyses since it was present within a well characterised sequence and was predicted to fold into a tRNA secondary structure (see below).

The Ruka sequences in *R. appendiculatus* and *B. (R.) microplus* were found to have the characteristics of a SINE. Direct repeats, 10–14 bp in length, were found flanking the sequence. The 5′ end of Ruka contains the A and B boxes of the Pol III promoter and there was also a tRNA-like region at the 5′ end. The 3′ end contained poly-pyrimidine tracts and was relatively A-rich compared to the rest of the Ruka sequence. Most, but not all of the, Ruka sequences were found to have a Polymerase III terminator, TTTT, at the 3′ end. The secondary structure prediction of the 5′ end of Ruka using tRNA Scan-SE (Lowe and Eddy, 1997) demonstrated that the sequence is capable of adopting a pseudo-tRNA structure of a serine tRNA. The amino acid that a given tRNA binds to is determined by its anticodon. Ruka, when folded into a tRNA structure, was predicted to have an anticodon with sequence GCU (indicated by a box in Fig. 3, which recognises and incorporates the amino acid serine. By contrast a lysine tRNA, which is more typical in SINEs, should contain a UUU or UUC anticodon. The *B. (R.) microplus* genomic sequence present in the intron of G6PDH

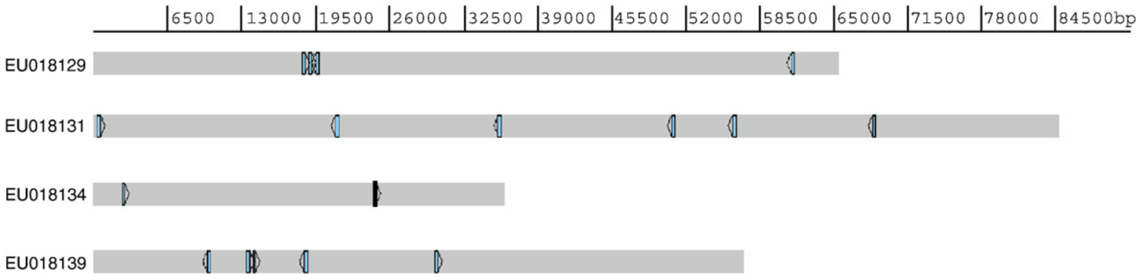

Fig. 1. Location of Ruka-SINEs within four *R. appendiculatus* contigs assembled from BAC genomic clone sequences. The scale in number of base pairs is at the top of the figure. Four contigs assembled from three different BAC clones – EU018129, EU018131, EU018134 and EU018139 – are shown as long grey rectangles. Open boxes within the sequence indicate the position of Ruka element with the arrows indicating the orientation of the SINE within the sequence.

```
                                                                 A
                                                          -----------
  1 CAGGTCACTTGCGTCTAGTGGCCCCGCCGCGGTGGTCTAGTGGCTAAGGTACTCGGCTGC
                  B
             ------------
 61 TGACCCGCAGGTCGCGGGTTCAAATCCCGGCTGCGGCGGCTGCATTTCTGATGGAGGCGG

121 AAATGTTGTAGGCCCGTGTACTCAGATTTGGGTGCACGTTAAAGAACCCTAGGTGGTCGA

181 AAtttccGGAGccctccACTACGGCGtctctcATAATCATCAGTGGttttGGGACGTTAA

241 AccccACATATGAATCAATCAATTGCGTCTAGTGGTACAAGTTTCGAG
```

Fig. 2. Ruka sequence from *R. appendiculatus*. Sequence of a Ruka-SINE located within a transcript within the *R. appendiculatus* gene index (RaGI). The Polymerase III promoter boxes A and B are indicated by dashed lines; a tRNA-related sequence is underlined; poly-pyrimidine tracts are in bold lower case text; short direct repeats are indicated by bold capitalised text.

was used to search the *R. appendiculatus* salivary gland EST gene index RaGI (Nene et al., 2004) using BLASTN (low complexity filter disabled) and produced 49 matches with an *E*-value less than 1e^{-10}; of which 34 were 180 bases or longer (Table 2). The sequence properties and tRNA secondary structure predictions of the matching sequences in the *R. appendiculatus* were similar to those of the genomic Ruka from *B. (R.) microplus*. In particular, the second Pol III motif (GGGTTCGATCCC for *R. appendiculatus*) was well conserved in sequence and position. Eleven random Ruka-containing clones from the RaGI cDNA library were sequenced, using vector primers and Ruka-specific internal primers. In ten of the 11 clones it was confirmed that the sequences were poly-adenylated (data not shown) demonstrating that these were genuine transcripts and not the result of genomic DNA contamination of the library. One clone could not be successfully sequenced hence presence of a polyA tail in its transcript could not be ascertained. Representative examples of a Ruka sequence and the predicted pseudo-tRNA secondary structure present in RaGI are shown in Figs. 2 and 3 respectively.

3.2. Presence of Ruka in EST databases derived from *B. (R.) microplus, A. variegatum and I. scapularis*

The gene indices for *B. (R.) microplus* (BmiGI), *A. variegatum* (AvGI) and *I. scapularis* (IsGI) were searched using the Ruka sequence from the G6PDH intron of *B. (R.) microplus* to identify sequences related to this element. Similar sequences with an *E*-value of ≤1e^{-10} were found in all the databases (Table 2).

The number of Ruka-like sequences of length 180 bases or more and with >80% similarity was >30 within both the BmiGI and RaGI databases. The AvGI database contained seven sequences with >80% identity to Ruka and additionally three sequences in IsGI exhibited approximately 70% identity. The Ruka sequences, longer than 180 bp, identified in the Ra BAC assemblies, RaGI and BmiGI, were aligned and an unrooted cladogram, shown in Fig. 4, was calculated using a tree-searching maximum likelihood algorithm. The tree contained two major clusters. One predominantly comprised sequences from *B. microplus*, whereas the second was a mixture of sequences from both *B. microplus* and *R. appendiculatus*. When the analysis was repeated using *I. scapularis* as an outgroup the results remained similar (data not shown). We presume that given the very close phylogenetic relationship of *Rhipicephalus* and *Boophilus (Rhipicephalus)*, which have recently been re-classified in the same genus (Murrell and Barker, 2004) the Ruka sequences in cluster two may have been present in the common ancestor of these two tick species.

3.3. Quantification of Ruka in the R. appendiculatus genome using real time PCR

Six primer pairs (described in Section 2.3.3 in Materials and methods and listed in Table 1) derived from the eight most conserved Ruka copies in the Ra BAC sequences were used to amplify Ruka sequences from the genomic DNA of *R. appendiculatus* using real time PCR. The results are summarised in Table 3. The genomic copy number estimates computed by the RT-PCR software following amplification using the six Ruka primer pairs were between 5100 and 28,800

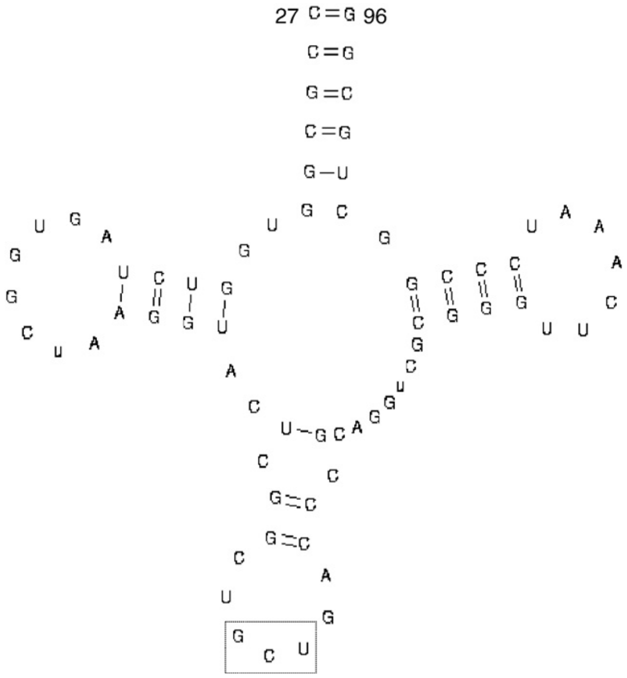

Fig. 3. Predicted secondary structure of Ruka. Secondary structure of the tRNA-related region of a SINE in a RaGI transcript as predicted by tRNA Scan-SE (Lowe and Eddy 1997). Numbers correspond to base positions underlined in Fig. 2. The boxed region is the serine anticodon.

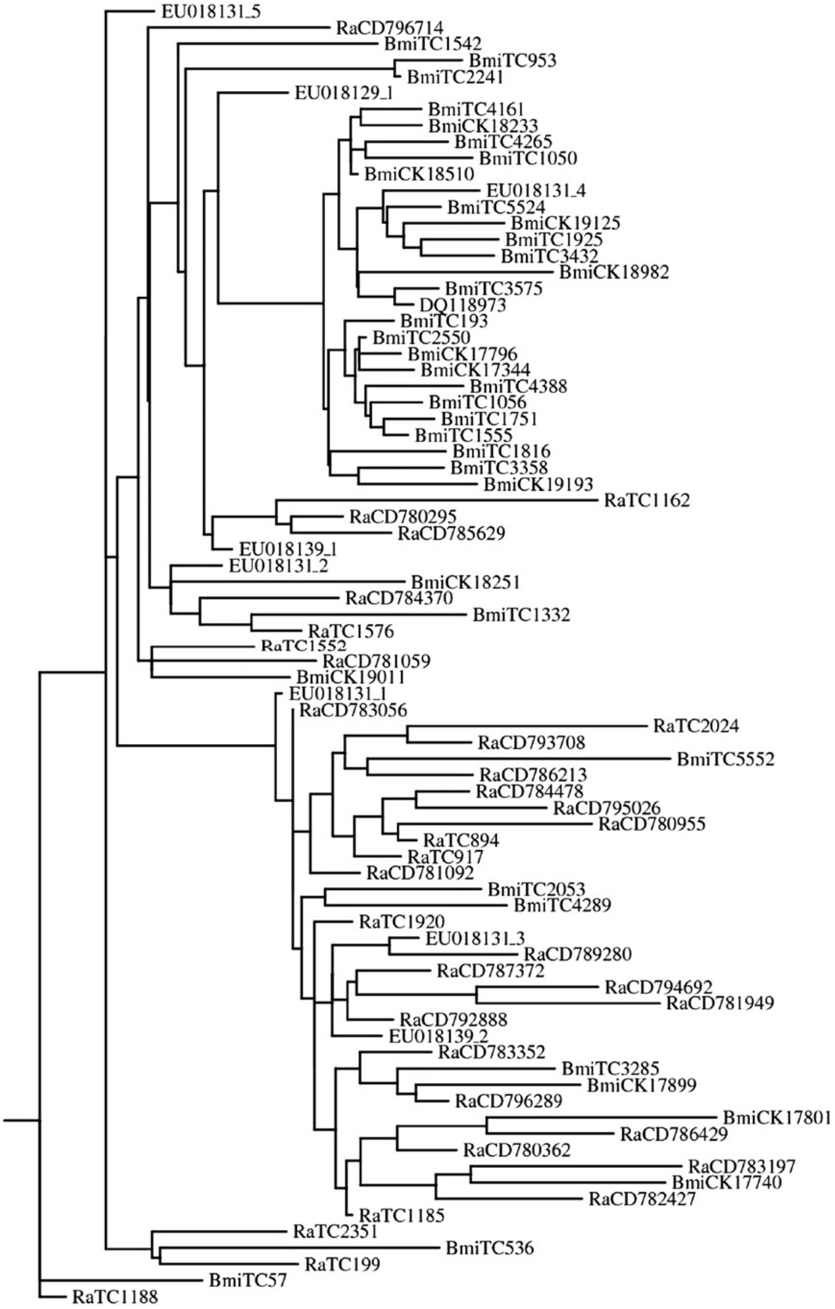

Fig. 4. Unrooted Cladogram illustrating the relatedness of 80 Ruka-like sequences within *B. (R.) microplus* (prefix Bmi) and *R. appendiculatus* (prefix Ra) transcripts and *R. appendiculatus* genomic sequences derived from genomic BAC clones (Accession numbers: EU018129, EU018131, EU018139). The cladogram was generated using a maximum likelihood algorithm after calculating 100 bootstrap trees by parsimony.

(Table 3, column 4). In order to provide an indication of the probability of a Ruka primer pair amplifying a specific copy of Ruka, the percentage similarity of the Ruka primer sequences to all the copies of Ruka in the Ra BAC sequences was calculated. A good match was defined as 100% identity in two or more bases at the 3′ end, combined with the entire primer sequence exhibiting >80% similarity with the relevant section of the cloned genomic copy (Table 3, column 5). This revealed that between 75% and 88% of the Ruka copies in these clones would theoretically be amplified by individual Ruka primer pairs 1–6. In addition to the lack of sequence conservation in some instances, the likelihood of primer competition for multiple copies of Ruka in genomic DNA strongly suggests that the copy numbers indicated by the quantitative-RT-PCR (qRT-PCR) results were almost certainly minimal estimates.

3.4. Evidence for additional categories of retroposons in R. appendiculatus genomic DNA and salivary gland transcripts

Additional types of transposable elements that have the capacity for autonomous retroposition, exploiting proteins encoded within their open reading frames have been described.

Table 3
RT-PCR using *R. appendiculatus* genomic DNA as a template with RUKA primer pairs 1–6

Primer pair	R^2	PCR efficiency	Average copy no. per haploid genome	% Ra BAC Ruka amplifiable	Predicted copy no. in haploid genome
Ruka 1	0.997	0.67	1.05E+04	75	1.40E+04
Ruka 2	0.997	0.78	5.11E+03	88	5.81E+03
Ruka 3	0.993	0.92	2.02E+04	88	2.30E+04
Ruka 4	0.992	0.74	2.88E+04	75	3.84E+04
Ruka 5	0.991	0.84	1.45E+04	88	1.65E+04
Ruka 6	0.995	0.88	1.14E+04	88	1.30E+04

The R^2 value (column 2) is the coefficient that is used to assess the fit of the standard curve to the data points plotted. The R^2 value was >0.99 which is the required value for reliable quantitation. The efficiency of the PCR reaction (E) (column 3) is calculated using the formula $E=(10^{(-1/\mathrm{slope})}-1)$, where the slope is calculated from a standard curve plot of Ct values against the logarithm of template amount. A value close to 1 indicates high PCR efficiency. Column 6 shows the modified predicted copy numbers following correction for the degree of sequence conservation between Ruka primer pairs 1–6 and the Ruka sequences within the BAC clones.

Domains within proteins, which enable independent transposition of retroposons, were used to interrogate the database using TBLASTX. The search revealed significant matches to constituents of reverse transcriptases, endonucleases and integrases, which comprise the transpositional apparatus of Class I retrotransposons, in a wide range of organisms. These matches had high levels of identity with both vertebrate and invertebrate genes (Table 4). When *R. appendiculatus* BAC sequences were searched against the databases using similar methodology, a putative long interspersed element (LINE) was identified (Accession no. EU018129). This element is approximately 7 kb in length and contains sequences that are predicted to encode reverse transcriptases and endonucleases. Sequence that encodes Class I transposable element-associated-proteins, including integrase, zinc-finger, retroviral protease and reverse transcriptase domains were also identified. Class II transposable elements that transpose by DNA duplication and insertion are frequently less abundant and more unevenly distributed within genomes than Class I retroposons (Wong and Choo, 2004; Kidwell, 2002). We were unable to detect any class II transposons through initial BLAST analysis of the *R. appendiculatus* sequence data. None of the repeat sequences appear to be similar to the R2 elements that have previously been described in ticks (Bunikis and Barbour, 2005). The database searches strongly suggest that additional retroposons are present in the *R. appendiculatus* genome.

3.5. Conservation of Ruka and the insertion sites

SINE insertion appears to be both random and irreversible, and therefore can be used to examine phylogenetic relationships through comparative analysis of insertion sites between different isolates, strains and species. The 20 tentative consensus sequences and singletons that contained the most conserved Ruka-like sequences identified within RaGI were queried against the BmiGI, AvGI and IsGI databases in order to detect sequence similarity within the regions flanking Ruka insertions. The same process was carried out for the 20 most

Table 4
Protein domains associated with transposable elements detected in RaGI using TBLASTX

RaGI no.	Protein similarity from non-redundant database	Protein database hit range/bp	*E*-value	Identity	Positives
TC1721	Similar to gag-pol polyprotein [*Strongylocentrotus purpuratus*] (XR_026136)	2877–3542	8E−117	139/219 (63%)	173/219 (78%)
CD795596	Similar to transposase (LOC575495) [*Strongylocentrotus purpuratus*] (XM_001191245)	1252–2133	1E−89	176/296 (59%)	221/296 (74%)
CD791315	Similar to LReO_3 [*Danio rerio*] (XM_001341375)	3058–3909	2E−75	138/304 (45%)	205/304 (67%)
CD792017	*Danio rerio* similar to novel transposon (XM_001338021)	2659–3345	2E−65	110/229 (48%)	167/260 (64%)
CD791975	Protease, reverse transcriptase, ribonuclease H, integrase [*Drosophila buzzatii*] (AJ133521)	4522–4932	7E−44	50/137 (36%)	133/238 (55%)
CD780002	novel transposon [*Danio rerio*] (XM_001343848)	1555–1842	8E−42	44/96 (45%)	124/197 (62%
CD794758	Similar to endonuclease/reverse transcriptase [*Strongylocentrotus purpuratus*] (XM_001197163)	2473–2931	4E−39	85/230 (36%)	134/230 (58%)
CD782785	Similar to endonuclease/reverse transcriptase [*Strongylocentrotus purpuratus*] (XM_001199602)	2458–3012	1E−37	81/185 (43%)	105/198 (53%)
CD793264	Similar to Gypsy polyprotein [*Danio rerio*] (XR_030000)	1396–1623	3E−36	33/76 (43%)	139/270 (51%)
CD791187	Biomphalaria glabrata “Line-like” repeat (X60372)	737–1354	1E−34	92/206 (44%)	143/206 (69%)
CD781634	Similar to endonuclease/reverse transcriptase [*Strongylocentrotus purpuratus*] (XM_001189330)	1765–2139	2E−31	37/125 (29%)	66/125 (52%)
CD780777	novel transposon [*Danio rerio*] (XM_001343848)	3055–3474	7E−25	54/140 (38%)	83/140 (59%)
CD796146	Similar to endonuclease/reverse transcriptase [*Strongylocentrotus purpuratus*] (XM_001194428)	883–1038	2E−18	18/52 (34%)	34/52 (65%)
CD782737	*Danio rerio* similar to novel transposon (LOC100008180) (XM_001346437)	481–726	7E−18	35/82 (42%)	52/82 (63%)

Table 5
Conserved Ruka insertions at common loci shared between two or more tick species

Locus	*R. appendiculatus*	*B. (R.) microplus*	*A. variegatum*	*I. scapularis*
1	CD787372	TC1925	BM292626	No match
2	TC1188	TC57	No match	No match
3	CD789280	TC3432	BM290387	No match
4	TC1552	CK191250	No match	No match

The identifier for the tentative consensus sequence (TC) and/or singleton (CD/CK/BM) is provided for each species.

similar sequences from the BmiGI and all the Ruka-like sequences (*E*-value ≤ $1e^{-10}$) from the AvGI and IsGI. Four loci with sequence similarity in the regions flanking the Ruka insertions were identified (Table 5 and Supplementary material Fig. 1) in *R. appendiculatus* and *B. (R.) microplus*. Two loci were common to three tick species, *R. appendiculatus*, *B. (R.) microplus* and *A. variegatum*. One of the Ruka insertion loci that was conserved between *B. (R.) microplus* and *R. appendiculatus*, but not in *A. variegatum* had the same tentative annotation (nucleoside di-phosphate kinase) in both species. These loci are likely to be the sites of ancestral insertions that occurred prior to the evolution of new ixodid genera and speciation within the genus *Rhipicephalus* (*Boophilus*). There was no convincing evidence for conservation of any of these loci in the *I. scapularis* gene index.

4. Discussion

We describe a family of tRNA-derived SINEs in ixodid ticks, that we have designated Ruka. If the frequency of occurrence of this SINE in selected BAC-derived genomic sequences is representative of the entire *R. appendiculatus* genome and assuming by analogy with data from three closely related ixodid species (Ullmann et al., 2005; Palmer et al., 1994) that the *R. appendiculatus* genome is ≥ 10^9 bp in size, there would be 65,000 copies of Ruka within the *R. appendiculatus* genome. Real time PCR analyses on genomic sequence using six distinct sets of primers that would each potentially amplify a subset of the total Ruka complement in *R. appendiculatus*, suggested that specific subsets of Ruka had genomic copy numbers in the range of approximately 5000–38,000 and that the overall total of Ruka-like SINES would probably be in excess of this figure. Due to uncertainty of the efficiency of which a specific primer pair would amplify multiple variant copies of Ruka in the *R. appendiculatus* genome it was difficult to provide a more precise estimate using this methodology. Given these factors, the agreement of the qRT-PCR data with the copy number extrapolation based on limited genomic sample sequencing is surprisingly good.

According to the anticodon sequence emerging from fold predictions the Ruka elements may be derived from a serine, rather than a lysine tRNA, although the latter is typically the most frequent progenitor of eukaryotic SINEs (Shedlock and Okada, 2000). There is still no confirmed explanation for the generally higher prevalence of lysine tRNA-derived SINEs observed in most organisms, but it is thought that this structure binds more effectively to reverse transcriptase and may therefore multiply more efficiently (Okada, 1991).

Ruka was found in all four species of ixodid tick analysed and is widespread in transcribed sequences in *B. (R.) microplus* and *R. appendiculatus*. It is interesting to note that most of the Ruka-containing transcripts do not contain long open reading frames (data not shown), suggesting widespread transcription of non-protein encoding DNA sequences in these two tick species. Ruka transcripts were generated by RNA Pol II, since the libraries from which the EST gene indices were constructed were derived from polyadenylated transcripts and Pol III transcripts are rarely polyadenylated (Duvel et al., 2003). The original method of priming for EST gene index construction depended on polyadenylation of the sequences and we observed polyA tails in 10 out of 11 randomly chosen Ruka-containing clones that were sequenced. The Ruka sequences found in RaGI and BmiGI are therefore very unlikely to be intermediates in the transposition process but are more likely indicative of promiscuous Pol II transcription. Another possible explanation is intron retention in Pol II transcripts, although there is no evidence that the Ruka copies observed in RaGI or BmiGI are located in introns.

SINEs are non-autonomous transposable elements and therefore do not encode the machinery required for transcription and insertion into the genome. Movement of tRNA-derived SINEs is mediated by proteins encoded within LINEs. It is thought that in most categories of SINEs, each element has a specific LINE partner with which it shares similarity in the 3′ end sequence. The common 3′ ends are the putative binding regions for the retroposition machinery (Okada et al., 1997). On the other hand, mammalian L1 and Alu SINEs do not share 3′ homology with LINEs. We did not observe any LINE sequences in the gene indices; but a LINE was identified in one of the genomic contigs derived from the Ra BAC clone sequences. This LINE does not appear to be the partner of Ruka as there is no sequence similarity between their respective 3′ ends. It is possible that Ruka and its corresponding LINE represent an exception to the general rule of shared 3′ identity, as in the case of the L1 and Alu SINEs. This suggests that if the SINEs that we identified are mobile, the genes required for their transposition are not encoded in the closely adjacent regions of the genome. A direct comparison of DNA from a different *R. appendiculatus* isolate would determine if the Ruka elements that we have discovered are actively mobile. A comparison of RaGI and BmiGI indicated that most insertions appear to be unique in their location. These indices are not equivalent in that RaGI is derived specifically from salivary gland tissues and is not normalised (Nene et al., 2004), whereas BmiGI is derived from a range of tissues and life cycle stages and is normalised (Guerrero et al., 2005). However, the conservation of the Pol III promoter sites between Ruka in *R. appendiculatus* and *B. (R.) microplus*, and also the identification of protein domains required for retroposition elsewhere in the genome of *R. appendiculatus* is consistent with mobility of these elements. No Class II DNA transposons that utilise a DNA replication mechanism as the basis of their mobility were identified in the subset of *R. appendiculatus* genomic or transcribed sequences that we analysed. However Class II DNA transposons do occur in insects including *Drosophila* (Kidwell, 2002). The failure to detect Class II transposons in this study may indicate, either that they are less abundant than Class I retroposons in tick genomes, or as in other arthropods including *Drosophila*, are localised in genomic regions such as heterochromatin that are under-represented in transcript databases (Wong and Choo, 2004).

There were four occurrences of common Ruka insertion flanking sequences in the genomes of *R. appendiculatus* and *B. (R.) microplus* as compared to two conserved flanking sites within the *A. variegatum* genome (Table 5). The result is consistent with phylogenetic analyses based on small subunit rRNA, ribosomal internal transcribed spacer and mitochondrial genome sequences (Murrell and Barker, 2004) indicating that *Amblyomma* is relatively divergent from *Rhipicephalus* and *Boophilus (Rhipicephalus)*. The shared insertion sites in RaGI and BmiGI that are absent from *Amblyomma* may well result from transposition of Ruka subsequent to the evolutionary separation of the two genera. The lack of common insertion loci within the IsGI is not surprising since the prostriate lineage, which contains *I. scapularis*, is highly divergent from the other three tick species that are classified within the metastriate lineage, according to molecular phylogeny, although there is no well defined evolutionary timescale for this split (Barker and Murrell, 2004, Fig. 1). There are major differences in the two lineages, including the overall chromosomal complement and the distribution of X chromosomes between males and females. Given the molecular divergence of the *I. scapularis* clade (Barker and Murrell, 2004) some elements within the ancestral Ruka transposable element family are likely to have undergone significant changes in *I. scapularis* resulting in the creation of novel SINEs that cannot be detected using sequence similarity-based comparisons, although secondary structure analyses may reveal additional copies. However, copies of Ruka were nonetheless observed in both *I. scapularis* (three copies) and *A. variegatum* (10 copies*)*. These SINEs may represent master copies that are conserved in sequence due to their frequent transposition (Kidwell, 2002).

Reassociation kinetic studies of tick genomes have shown that they are rich in both highly repetitive and moderately repetitive DNA. The data presented herein suggest that the high percentage of the moderately repetitive DNA detected in ixodid genomes by reassociation kinetics (Palmer et al., 1994; Ullmann et al., 2005) can partially be explained by the frequent occurrence of transposable elements of the class I group that transpose via RNA intermediates. As the current *I. scapularis* genome sequencing project continues, a more comprehensive dataset will become available to test this hypothesis.

Acknowledgements

We are grateful for the skilled assistance of Stephen Mwaura and other members of the ILRI tick unit. We thank Drs. Morris Agaba and Philip Toye for their helpful comments, and appreciate the assistance of Dr. David Herndon in the preparation of BAC DNA and sequencing of BAC ends. We thank USAID and the Extramural Program of USDA-ARS for funding this work. This is ILRI publication number 200704.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.gene.2008.01.026.

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410.

Applied Biosystems, 2003. (Available: www.appliedbiosystems.com/support/tutorials/pdf/quant_pcr.pdf) [Accessed November 2007].

Barker, S.C., Murrell, A., 2004. Systematics and evolution of ticks with a list of valid genus and species nwww.appliedbiosystems.com/support/tutorials/pdf/quant_pcr.pdfames. Parasitology 129, S15–S36 Suppl.

Benson, G., 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 27, 573–580.

Betley, N.J., Frith, M.C., Graber, J.H., Choo, S., Deshler, J.O., 2002. A ubiquitous and conserved signal for RNA localization in chordates. Curr. Biol. 12, 1756–1761.

Bunikis, J., Barbour, A.G., 2005. Ticks have R2 retrotransposons but not the consensus transposon target site of other arthropods. Ins. Mol. Biol. 14, 465–474.

Chenna, R., et al., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3497–3500.

Clamp, M., James, C., Searle, S., Barton, G., 2004. The Jalview Java alignment editor. Bioinformatics 20, 426–427.

De Castro, J., 1997. Sustainable tick and tick-borne disease control in livestock improvement in developing countries. Vet. Parasitol. 17, 77–97.

Desjardins, C., et al., 2007. Structure and evolution of a proviral locus of *Glyptapanteles indiensis* bracovirus. BMC Microbiol 7, 61–78.

Duvel, K., Pries, R., Braus, G.H., 2003. Polyadenylation of rRNA and tRNA-based yeast transcripts cleaved by ribozyme activity. Curr. Genet. 43, 255–262.

Felsenstein, J., 1989. PHYLIP — Phylogeny Inference Package (Version 3.2). Cladistics 5, 164–166.

Geraci, N.S., Johnston, J.S., Robinson, J.P., Wikel, S.K., Hill, C.A., 2007. Variation in genome size of argasid and ixodid ticks. Insect Biochem. Mol. Biol. 37, 399–408.

Godornes, C., Leader, B.T., Molini, B.J., Centurion-Lara, A., Lukehart, S.A., 2007. Quantitation of Rabbit Cytokine mRNA by Real-Time RT-PCR. Cytokine 38, 1–7.

Guerrero, F.D., et al., 2005. BmiGI: a database of cDNAs expressed in *Boophilus microplus*, the tropical/southern cattle tick. Insect Biochem. Mol. Biol. 35, 585–595.

Jongejan, F., Uilenberg, G., 2004. The global importance of ticks. Parasitol. 129, S3–S14.

Junier, T., Pagni, M., 2000. Dotlet: diagonal plots in a web browser. Bioinformatics 16, 178–179.

Kaminker, J.S., et al., 2002. The transposable elements of the *Drosophila melanogaster* euchromatin: a genomics perspective. Genome Biol. 3 RESEARCH0084.

Kidwell, M.G., 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115, 49–63.

Klompen, J.S.H., Black IV, W.C., Keirans, J.E., Oliver Jr, J.H., 1996. Evolution of ticks. Annu. Rev. Entomol. 41, 141–161.

Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25, 955–964.

Murrell, A., Barker, S.C., 2004. Systematics and evolution of ticks with a list of valid genus and species names. Parasitology 129, S15–S36.

Nene, V., et al., 2002. AvGI, an index of genes transcribed in the salivary glands of the ixodid tick *Amblyomma variegatum*. Int. J. Parasitol. 32, 1447–1456.

Nene, V., et al., 2004. Genes transcribed in the salivary glands of female *Rhipicephalus appendiculatus* ticks infected with *Theileria parva*. Insect Biochem. Mol. Biol. 34, 1117–1128.

Okada, N., 1991. SINEs: short interspersed repeated elements of the eukaryotic genome. Trends Ecol. Evol. 6, 358–361.

Okada, N., Hamada, M., Ogiwara, I., Ohshima, K., 1997. SINEs and LINEs share common 3′ sequences: a review. Gene, 205, 229–243.

Palmer, M.J., Bantle, J.A., Guo, X., Fargo, W.S., 1994. Genome size and organization in the ixodid tick *Amblyomma americanum (L.)*. Insect Mol. Biol. 3, 57–62.

Petrov, D.A., Sangster, T.A., Johnston, J.S., Hartl, D.L., Shaw, K.L., 2000. Evidence for DNA loss as a determinant of genome size. Science 287 (5455), 1060–1062.

Quackenbush, J., Liang, F., Holt, I., Pertea, G., Upton, J., 2000. The TIGR gene indices: reconstruction and representation of expressed gene sequences. Nucleic Acids Res. 28, 141–145.

Quackenbush, J., et al., 2001. The TIGR gene indices: analysis of gene transcript sequences in highly sampled eukaryotic species. Nucleic Acids Res. 29, 159–164.

Ribeiro, J.M.C., et al., 2006. An annotated catalog of salivary gland transcripts from *Ixodes scapularis* ticks. Insect Biochem. Mol. Biol. 36, 111–129.

Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. Molecular Cloning: A Laboratory Manual, 2nd ed. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Shedlock, A.M., Okada, N., 2000. SINE insertions: powerful tools for molecular systematics. BioEssays 22, 148–160.

Smith, T.F., Waterman, M.S., 1981. Identification of common molecular subsequences. J. Mol. Biol. 147, 195–197.

Sobreira, T.J.P., Durham, A.M., Gruber, A., 2006. TRAP: automated classification, quantification and annotation of tandemly repeated sequences. Bioinformatics 22, 361–362.

The Gene Index Databases, Dana Farber Cancer Institute, Boston, MA 02115 (Available: http://biocomp.dfci.harvard.edu/tgi/tgipage.html) [Accessed July 2007].

Tu, Z., 1999. Genomic and evolutionary analysis of *Feilai,* a diverse family of highly reiterated SINEs in the yellow fever mosquito. Mol. Biol. Evol. 16, 760–772.

Tu, Z., Li, S., Mao, C., 2004. The changing tails of a novel short interspersed element in *Ædes ægypti*. Genetics 168, 2037–2047.

Ullmann, A.J., Lima, C.M.R., Guerrero, F.D., Piesman, J., Black, W.C., 2005. Genome size and organization in the blacklegged tick, *Ixodes scapularis* and the Southern cattle tick, *Boophilus microplus*. Insect Mol. Biol. 14, 217–222.

Ullu, E., Tschudi, C., 1984. Alu sequences are processed 7SL RNA genes. Nature 312, 171–172.

Wong, L.H., Choo, K.H.A., 2004. Evolutionary dynamics of transposable elements at the centromere. Trends Genet. 20, 611–616.